

***"This is the peer reviewed version of the following article: Kemppainen, P., Rønning, B., Kvalnes, T., Hagen, I. J., Ringsby, T. H., Billing, A. M., ... & Jensen, H. (2017). Controlling for P-value inflation in allele frequency change in experimental evolution and artificial selection experiments. Molecular ecology resources, 17(4), 770-782., which has been published in final form at 10.1111/1755-0998.12631. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving."***

1 **Title**

2 Controlling for  $p$ -value inflation in allele frequency change in experimental  
3 evolution and artificial selection experiments.

4

5 **Authors**

6 Petri Kemppainen\*, Bernt Rønning\*, Thomas Kvalnes\*, Ingerid J. Hagen\*, Thor-  
7 Harald Ringsby\*, Anna M. Billing\*, Henrik Pärn\*, Sigbjørn Lien†, Arild Husby\*‡,  
8 Bernt-Erik Sæther\* and Henrik Jensen\*

9

10 **Affiliations**

11 \*Centre for Biodiversity Dynamics, Department of Biology, Norwegian University  
12 of Science and Technology, NO-7491 Trondheim, Norway.

13 † CIGENE, Norwegian University of Life Sciences, P.O. Box 5003, NO-1432 Ås,  
14 Norway.

15 ‡ Department of Biosciences, P.O. Box 65 (Viikinkaari 1), 00014 University of  
16 Helsinki, Finland.

17

18 **Keywords**

19 Experimental evolution, artificial selection, population stratification,  $p$ -value

20 inflation, relatedness, genome wide association studies

21

22 **Corresponding author**

23 Petri Kemppainen, Centre for Biodiversity Dynamics, Department of Biology,

24 Norwegian University of Science and Technology, Høgskoleringen 5,

25 Realfagbygget E1-126, NO-7491, Trondheim, Norway.

26

27 Phone number: +4745394509

28 E-mail address: petrikemppainen2@gmail.com

29

30 **Running title**

31  $P$ -value inflation in experimental evolution

32 **Abstract**

33 Experimental evolution studies can be used to explore genomic response to  
34 artificial and natural selection. In such studies, loci that display larger allele  
35 frequency change than expected by genetic drift alone are assumed to be directly  
36 or indirectly associated with traits under selection. However, such studies report  
37 surprisingly many loci under selection, suggesting that current tests for allele  
38 frequency change may be subject to  $p$ -value inflation and hence be anti-  
39 conservative. One factor known from genome wide association (GWA) studies to  
40 cause  $p$ -value inflation is population stratification, such as relatedness among  
41 individuals. Here we suggest that by treating presence of an individual in a  
42 population after selection as a binary response variable, existing GWA methods  
43 can be used to account for relatedness when estimating allele frequency change.  
44 We show that accounting for relatedness like this effectively reduces false  
45 positives in tests for allele frequency change in simulated data with varying  
46 levels of population structure. However, once relatedness has been accounted  
47 for, the power to detect causal loci under selection is low. Finally, we  
48 demonstrate the presence of  $p$ -value inflation in allele frequency change in  
49 empirical data spanning multiple generations from an artificial selection  
50 experiment on tarsus length in two wild populations of house sparrow, and  
51 correct for this using genomic control. Our results indicate that since allele  
52 frequencies in large parts of the genome may change when selection acts on a  
53 heritable trait, such selection is likely to have considerable and immediate  
54 consequences for the eco-evolutionary dynamics of the affected populations.

55

56

57 **Introduction**

58 Phenotypic evolution experiments have been imperative for our understanding  
59 of both short and long-term evolutionary responses to selection (Dudley *et al.*  
60 1977; Palmer & Dingle 1986; Gromko *et al.* 1991; Hill & Caballero 1992; Gromko  
61 1995; Brakefield 2003; Conner 2003; Garland 2003). With increasing availability  
62 of population genomic data, it has become feasible to target the genomic changes  
63 that underlie phenotypic changes in such experiments (Ellegren & Sheldon 2008;  
64 Pardo-Diaz *et al.* 2015; Schlötterer *et al.* 2015). Two approaches that can be used  
65 to study genomic responses of selection are; (1) artificial selection, where  
66 individual survival or reproduction is artificially manipulated based on traits of  
67 interest (Heidaritabar *et al.* 2014) and (2) natural selection experiments, where  
68 survival and reproduction instead depends on the individuals inherent ability to  
69 cope with the environmental conditions (laboratory or natural) they are  
70 subjected to (Burke *et al.* 2010; Zhou *et al.* 2011; Turner *et al.* 2011; Remolina *et*  
71 *al.* 2012; Pespeni *et al.* 2013; Tobler *et al.* 2014; Gompert *et al.* 2014; Schlötterer  
72 *et al.* 2015). These studies often assume that loci showing significant allele  
73 frequency change following an episode of selection (e.g. when observed change  
74 falls outside the 95% quantiles of an appropriate null-distribution) are  
75 associated with the trait under selection (Barrett & Hoekstra 2011; Pespeni *et al.*  
76 2013; Gompert *et al.* 2014; Heidaritabar *et al.* 2014). Such associations can stem  
77 from loci directly affecting the trait under selection, or indirectly through genetic  
78 correlations deriving from linkage disequilibrium (LD; Nielsen 2005; Barrett &  
79 Hoekstra 2011). Studies of allele frequency change following episodes of  
80 selection like this are valuable because they can give insights into both the

81 number and the type of genes associated with potentially highly complex  
82 adaptations.

83         Genome wide association (GWA) studies are powerful tools to dissect the  
84 genetic architecture of quantitative and binary traits (McCarthy *et al.* 2008; Bush  
85 & Moore 2012). In such studies, it is widely recognized that relatedness at any  
86 level of the population hierarchy, ranging from family structure to population  
87 structure at different spatial scales (here collectively referred to as population  
88 stratification) may cause long range LD between loci (Korte & Farlow 2013). In  
89 turn, this may lead to false association between genotypes and phenotypes, often  
90 evident as substantial  $p$ -value inflation and large numbers of false positives  
91 (Devlin & Roeder 1999; Devlin *et al.* 2001; Marchini *et al.* 2004; Price *et al.*  
92 2010). As in GWA studies, test statistics for allele frequency change in  
93 experimental evolution rely on associations between genotypes and phenotypes.  
94 However, the possibility of  $p$ -value inflation due to population stratification in  
95 tests for allele frequency change have repeatedly been overlooked (Burke *et al.*  
96 2010; Zhou *et al.* 2011; Turner *et al.* 2011; Turner & Miller 2012; Remolina *et al.*  
97 2012; Pespenti *et al.* 2013; Turner *et al.* 2013; Gompert *et al.* 2014; Heidaritabar  
98 *et al.* 2014). These studies have consequently identified a surprisingly large  
99 number of loci putatively under selection (i.e. candidate loci). These findings  
100 were first questioned by Tobler *et al.* (2014), who showed that most of the  
101 identified candidate SNPs indeed were false positives, both by replicated  
102 experiments in *Drosophila melanogaster*, and in simulations. The false positives  
103 were mainly attributed to long range LD; either occurring naturally in the  
104 population (due to undetected population stratification) or as a consequence of  
105 the founders in the experiment representing only a small sample of the much

106 larger natural population. The mechanisms that cause  $p$ -value inflation in GWA  
107 studies are potentially the same that cause  $p$ -value inflation in allele frequency  
108 change in experimental evolution. While showing the potential for  $p$ -value  
109 inflation, Tobler *et al.* (2014) did not suggest any approaches to estimate its  
110 magnitude or to adjust for it. Here we demonstrate how methods already  
111 available to account for  $p$ -value inflation in GWA studies can be applied to  
112 genomic data from experimental evolution studies as well.

113         An appealing approach to study the effects of selection on genome  
114 variation is to estimate the population mean allele frequency change before and  
115 after selection (Pespeni *et al.* 2013; Gompert *et al.* 2014). If these episodes of  
116 selection occur within a single generation, the effects of drift and selection on  
117 such allele frequency change (estimated separately for each individual locus) are  
118 isolated from other processes, such as recombination and mutation, and  
119 empirical null-distributions can be generated by random permutation of samples  
120 (Pespeni *et al.* 2013; Gompert *et al.* 2014). As random permutation of samples  
121 does not take into account relatedness between individuals, we here  
122 demonstrate with simulations that estimating significance of allele frequency  
123 change like this is highly susceptible to  $p$ -value inflation arising from population  
124 stratification. As a means to account for  $p$ -value inflation, we propose that allele  
125 frequency change before and after selection can be tested using binary GWA  
126 analyses, where relatedness is included as a random effect (Aulchenko *et al.*  
127 2007). Such tests are applicable for data sets where samples of individuals are  
128 individually genotyped prior to a single episode of natural or artificial selection,  
129 and the same individuals can be classified as either present or absent in the  
130 population following the selection episode. Hence, we have here not considered

131 other types of data such as those from pooled sequencing experiments (e.g. Parts  
132 *et al.* 2011; Illingworth *et al.* 2012).

133           Whenever residual  $p$ -value inflation exists in the data, it is common  
134 practice in GWA studies to perform genomic control (GC; Price *et al.* 2010). The  
135 inflation factor ( $\lambda$ ) can be estimated by regression in a Q-Q plot, comparing  
136 observed versus expected (under the null-distribution) association statistics  
137 (Clayton *et al.* 2005), and GC is subsequently achieved by dividing the observed  
138 association statistics by  $\lambda$ . We test the merits of binary GWA analyses and GC on  
139 allele frequency change before and after selection using simulated population  
140 genomic data with varying levels of population structure. To demonstrate the  
141 close relationship between testing for allele frequency change in a GWA  
142 framework like this, and GWA analyses on the underlying quantitative trait  
143 under selection, we also compare results from the two different approaches,  
144 when relevant. The correlation between  $p$ -values from these two tests will give  
145 an indication to what extent they identify the same genomic regions being  
146 associated with the trait under selection.

147           Finally, as a demonstration of the concepts developed, we evaluate the  
148 occurrence of  $p$ -value inflation on empirical SNP data from an artificial selection  
149 experiment on two free-living island populations of house sparrow (*Passer*  
150 *domesticus*). In the experiment, tarsus length was artificially selected to increase  
151 or decrease across four consecutive years (2002-2005), resulting in an average  
152 phenotypic change of 0.5-0.6% per year in the expected directions (Kvalnes *et*  
153 *al.*, in review). Furthermore, it was shown that this change had a genetic basis:  
154 the average breeding values for tarsus length of cohorts produced on the two  
155 islands during these four years also changed in the directions predicted by the



156 artificial selection, and these changes were larger than expected due to genetic  
157 drift (Kvalnes et al., in review). Due to overlapping generations in the house  
158 sparrow (Jensen et al. 2008), allele frequency change over the whole  
159 experimental period cannot easily be tested directly with binary GWA analyses.  
160 Instead,  $p$ -values for allele frequency change were obtained from empirical null-  
161 distributions produced by gene-dropping simulations and represents thus a  
162 more complex study design compared to estimating allele frequency change  
163 within a single generation.

164

## 165 **Materials and Methods**

### 166 *Simulated population genomic data*

167 Simulated population genomic data sets were generated with the software  
168 *fastsimcoal2* (Excoffier & Foll 2011; Excoffier *et al.* 2013) with three  
169 chromosomes of 1Mb each, mutation rate of  $= 3 \times 10^{-8}$ , recombination rate of  
170  $1 \times 10^{-8}$  and no transition bias. With these parameters at least 5000 polymorphic  
171 SNPs were generated for all data sets. In data sets without population structure  
172 ('random mating') we set the effective population size ( $N_e$ ) to 20000. This is  
173 equivalent to two populations of  $N_e=10000$ , each exchanging half of the  
174 population as migrants each generation (i.e.  $N_e m = N_e/2$ , where  $m$  is the  
175 proportion of migrants exchanged each generation). In data sets with population  
176 structure we set the number of populations to two with  $N_e = 10000$  each and  $N_e m$   
177  $= 2$  ('moderate population structure') or  $N_e m = 1$  ('strong population structure').  
178 A relatively large  $N_e$  ensured that LD quickly declines with physical distance.  
179 From simulations with no population structure we sampled 100 diploid  
180 individuals and with population structure we sampled 50 diploid individuals

181 from each of the two populations. Five thousand bi-allelic SNPs with a minor  
182 allele frequency (MAF) above 0.05 were randomly chosen to create data sets of  
183 equal sizes for all levels of population structure.

184 For each replicate simulated data set, two, four or eight loci were  
185 randomly chosen to represent causal loci. For each causal locus, one allele was  
186 randomly chosen to translate to a phenotypic value of one with the alternative  
187 translating to a phenotypic value of zero, giving phenotypic values of 0, 1 or 2 for  
188 genotypes at each causal locus. The final phenotypic value for each individual  
189 was the sum of these values across the causal loci, with Gaussian noise added to  
190 generate a narrow sense heritability of  $h^2 = 0.5$ , defined as  $V_A/V_P$ , where  $V_P$  is the  
191 total phenotypic variance and  $V_A$  the additive genetic variance (which were  
192 known in our simulated data). Individuals with phenotypic values above the  
193 mean plus 0.3 standard deviations of the mean were considered as ‘surviving’  
194 corresponding to an average selection intensity of one (Falconer & Mackay  
195 1996). To simulate no heritability, phenotypic values were randomized among  
196 individuals prior to analyses. For each combination of levels in population  
197 structure and number of causal loci we generated 100 replicates, resulting in a  
198 total 900 of simulated data sets. In analyses with and without heritability the  
199 same simulated data sets were used.

200

### 201 *Linkage disequilibrium*

202 Linkage disequilibrium is well known to increase with population structure.  
203 Here we present analyses of LD of the simulated data sets mainly as a  
204 background for discussing its role in causing  $p$ -value inflation in tests for allele  
205 frequency change. Linkage disequilibrium was estimated as the coefficient of

206 determination between pairs of loci ( $r^2$ ) for all pairwise comparisons between  
207 500 randomly chosen SNPs from each simulated data set using the function  
208 *r2fast* from the R-package *GenABEL* (Hao *et al.* 2007; Aulchenko *et al.* 2007).  
209 Linkage disequilibrium was considered as short range when estimated between  
210 pairs of loci closer than 10 kilo base pairs (kbp) from each other (loci closer than  
211 1 % on a chromosome, or  $\sim 1$  centiMorgan [cM] as recombination rate was set to  
212 be constant along chromosomes). Linkage disequilibrium between loci on  
213 different chromosomes was used as a proxy for all long range LD. The decay of  
214 LD with physical distance was estimated following Hill and Weir (1988); a non-  
215 linear model was fitted between LD and distance in kbp and LD half decay  
216 distance was estimated as the distance at which LD is half of its predicted  
217 maximum value.

218

#### 219 *GWA analyses and allele frequency change following selection*

220 The association between allelic variants of loci and phenotype were tested in the  
221 R package *GenABEL* (Aulchenko *et al.* 2007). To account for relatedness, a  
222 kinship matrix, *K*, was estimated by the *ibs* function, which calculates the average  
223 identity by state (IBS) for all pairs of individuals. The function *polygenic* was  
224 used to estimate residual trait variance and the inverse of the variance-  
225 covariance matrix in the presence of relatedness. Outputs from function  
226 *polygenic* were further analyzed with function *mmscore*, which implements the  
227 score test for association between genetic polymorphisms and a trait (Chen &  
228 Abecasis 2007). The *mmscore* function can be used on both quantitative and  
229 binary traits, which allowed us to (1) test allele frequency change before and  
230 after selection by treating viability as a binary response variable (binary), and

231 (2) directly compare this to tests for associations between genotypes and the  
232 underlying quantitative phenotypic trait under selection (quantitative). For the  
233 binary GWA analyses in the simulated data, we coded all individuals with  
234 phenotypic values larger than the mean plus 0.3 standard deviations of the mean  
235 (see above) as '1', representing individuals present in the population after  
236 selection, else they were coded as '0' (not present in the population after  
237 selection). Note that in such analyses of selection experiments one assumes that  
238 selection acts on one or more unknown but heritable trait(s), and thus that the  
239 only 'phenotypic' information needed for each individual present before  
240 selection is its presence/absence in the population also after selection.

241 We also performed all analyses ignoring relatedness by setting all  
242 pairwise IBS values to zero. In the absence of covariates, this reduced our GWA  
243 analyses to linear regressions. This was done for two reasons; (1) it allowed us  
244 to estimate the  $p$ -value inflation caused by population stratification when  
245 relatedness is not taken into account, and (2) it allowed us to compare binary  
246 GWA analyses on viability to previously used permutation tests for assessing  
247 significance of allele frequency change before and after selection. For the  
248 permutation tests, empirical null-distributions for allele frequency change before  
249 and after selection were generated by random permutation of samples as in  
250 Gompert *et al.* (2014). To avoid unnecessary replication but still achieve  
251 reasonable precision of estimated  $p$ -values, we continued permutations until at  
252 least 10 permuted values were more extreme than the observed, with a  
253 minimum 1000 permutations for all tests. This approach is similar to a  
254 sequential probability ratio test (Fay *et al.* 2007). Due to the large number of  
255 permutations required by the above procedure, the comparisons between binary

256 GWA analyses (ignoring relatedness) and the permutation tests were restricted  
257 to four data sets for each combination of levels of population structure and  
258 number of causal loci (36 data sets in total).

259         The amount of residual  $p$ -value inflation due to population stratification  
260 was estimated by regression in a Q-Q plot based on observed versus expected  $\chi^2$ -  
261 values under the null-distribution. The inflation factor,  $\lambda$ , is the slope of the  
262 regression, where  $\lambda$ -values larger than one indicate  $p$ -value inflation. Although  
263 no strict guidelines exist, here we considered  $\lambda > 1.1$  to indicate strong  $p$ -value  
264 inflation.

265         In our simulated data, we tested the correlation of  $p$ -values ( $-\log_{10}$   
266 transformed) from the binary and quantitative GWA analyses. A strong  
267 correlation indicates that both tests identify the same genomic regions being  
268 associated with the trait under selection. To test the extent to which the  
269 underlying population structure (rather than true genetic correlations) in the  
270 data affects the outcomes of these tests more generally, we also tested the  
271 correlation of  $p$ -values from binary and quantitative GWA analyses when traits  
272 were based on two different sets of causal loci. To assure independence, we did  
273 not allow any of the causal loci from the two sets to be closer than 100 kbp from  
274 each other ( $\sim 10$  cM). If population stratification has no influence on  $p$ -values, the  
275 expected number of significant correlations from such tests should be close to  
276 5% and display no inflation (in a Q-Q plot comparing  $-\log_{10} p$ -values) compared  
277 to  $p$ -values following a uniform null-distribution. To investigate how associations  
278 between genotypes and phenotypes depends on population stratification, we  
279 tested to what extent the  $t$ -statistics from the  $p$ -value correlations between  
280 binary and quantitative GWA analyses in turn correlated with the mean of  $\log_{10} \lambda$

281 for each pair of tests. This value,  $\log_{10} \bar{\lambda}$ , was used as a proxy for how much  
282 population stratification there was in the data, which could vary considerably  
283 also within the different levels of population structure (note that population  
284 stratification can also be present in data sets with no population structure, see  
285 also Discussion). The same 900 simulated data sets as above were used except  
286 we dropped the number of causal loci as a factor and used four causal loci for all  
287 analyses (i.e.  $n = 300$  for each level of population structure).

288 To control for multiple testing, we estimated q-values (expected  
289 proportion of false positives among all tests that are deemed significant) using  
290 the function *qvalue* from Bioconductor's *qvalue* package (Dabney & Storey 2014).  
291 We considered a test significant when  $q < 0.1$ , i.e. accepting a 10% probability  
292 that that the test is a false positive. Here we define power of a test as the average  
293 number of significant causal loci, and all significant loci further than 50 kbp ( $\sim 5$   
294 cM) away from any causal loci were considered false positives. This distance is  
295 likely to be appropriate considering the distance at which LD breaks down in the  
296 simulated data set (see Results).

297

### 298 *Artificial selection in house sparrows and SNP genotyping*

299 An artificial selection experiment on tarsus length in two populations of house  
300 sparrows was conducted during the years 2002 to 2005 as described in Kvalnes  
301 et al. (in review). In short, for four successive years (2002-2005)  $\sim 90\%$  of all  
302 individuals on each of two islands (Leka and Vega) in northern Norway were  
303 captured each February, during approximately two weeks. At the end of this two-  
304 week period, all individuals with a tarsus longer than the mean plus 0.3 standard  
305 deviations of the mean were released back to Leka, and individuals with a tarsus

306 shorter than the mean minus 0.3 standard deviations of the mean were released  
307 back to Vega. These individuals comprised the selected individuals. The  
308 remaining individuals (non-selected) were relocated to distant mainland  
309 populations > 95 km's away. Thus, the strength of selection was the same as for  
310 the simulated data above. Individuals were genotyped at fourteen microsatellite  
311 loci to establish high quality genetic pedigrees (Rønning *et al.* 2016). Individuals  
312 with the most informative family links (File S1, Supporting Information) were  
313 chosen for genotyping on a custom house sparrow 10 K Illumina iSelect HD  
314 BeadChip (Hagen *et al.* 2013). Of the initial 10000 SNPs, 6492 were variable, of  
315 high quality and could be mapped to a reference genome (Hagen *et al.*, in  
316 preparation). This data was further filtered such that no more than 20% of  
317 genotypes were missing for any locus (median < 0.1%) or individual (median =  
318 0%). Loci that at some point (within an island) became fixed during the  
319 experimental period were ignored, as a null-distribution for such loci for those  
320 years cannot be generated. These procedures resulted in 5131 (from 267  
321 individuals) and 5075 SNPs (from 273 individuals) available for analysis on the  
322 island of Leka and on Vega, respectively. More detailed sample information is  
323 available in File S1 (Supporting Information).

324

### 325 *GWA analyses and allele frequency change in house sparrows*

326 GWA analyses on tarsus length were conducted on the two islands separately  
327 using the same data sets as used for testing allele frequency change. Because  
328 tarsus length does not change with age, we used mean values adjusted for  
329 fieldworker (Kvalnes *et al.* in review) when multiple measurements for adult

330 individuals were available (Jensen *et al.* 2003; 2008). For the function *polygenic*  
331 sex was included as fixed factor.

332 Allele frequency change was estimated within each island as the  
333 population mean allele frequency in all adult individuals immediately before  
334 artificial selection a given year (baseline), minus the population mean allele  
335 frequency in adult individuals present in the population directly after artificial  
336 selection (i.e. excluding the individuals that were removed from the island that  
337 year; see above). The total allele frequency change due to artificial selection for  
338 the experiment was attained by the sum of all the within-year changes. Thus, loci  
339 with large allele frequency changes in the same direction each year have the  
340 highest total allele frequency changes. Note that this only measures allele  
341 frequency change directly due to artificial selection and does not take into  
342 account the fact that drift and/or natural selection also may cause allele  
343 frequencies in the population to change between two successive artificial  
344 selection episodes. This was done to isolate the effect of the artificial selection on  
345 allele frequency change. *P*-values for allele frequency change for each locus were  
346 attained from an empirical null-distribution acquired from gene-dropping  
347 simulations (Gratten *et al.* 2012; File S2, Supporting Information). *P*-value  
348 inflation in gene-dropping simulations is likely to stem from the presence of  
349 relatedness among the founders; in the simulations founders are assumed only  
350 to be related by chance (File S2, Supporting Information). To correct for *p*-value  
351 inflation in the gene-dropping simulations, we performed GC by adjusting for  $\lambda$ ,  
352 which was estimated directly from  $-\log_{10} p$  (Price *et al.* 2010). Function *qvalue*  
353 was used to estimate *q*-values and the proportions of genes for which the null  
354 hypothesis is true ( $1-\pi_0$ ).



355

## 356 **Results**

### 357 *P-value inflation in allele frequency change before and after selection*

358 When relatedness was ignored in the binary GWA analyses, the correlations  
359 between  $-\log_{10} p$  from random permutation of samples and binary GWA analyses  
360 (both testing for allele frequency change before and after selection), were close  
361 to unity for all 36 simulated data sets (all  $r_p > 0.99$ ). There was no significant  
362 effect of population structure ( $P = 0.65$ ,  $F_{(2,27)} = 0.44$ ) or number of causal loci ( $P$   
363  $= 0.86$ ,  $F_{(2,27)} = 0.15$ ) on these correlations. Thus, when ignoring relatedness,  
364 binary GWA analyses can be considered as a proxy for previously used  
365 permutation tests for assessing significance of allele frequency change before  
366 and after selection.

367 For both GWA testing for allele frequency differences before and after  
368 selection with viability treated as binary response variable and GWA analyses  
369 performed on the underlying quantitative trait under selection, heritability is the  
370 main prerequisite for  $p$ -value inflation to occur (Fig. S1, Supporting Information).  
371 Thus, we present result on heritable traits only. When ignoring relatedness,  
372 considerable  $p$ -value inflation existed in data sets simulated under random  
373 mating (Fig. 1 A) for both binary and quantitative GWA analyses. This  $p$ -value  
374 inflation increased drastically with increasing population structure (Fig. 1 B).  
375 However, accounting for relatedness greatly reduced  $p$ -value inflation in all cases  
376 (Fig. 1 B).

377 False positive rates and power to detect causal loci for binary GWA  
378 testing for allele frequency change before and after selection reflect the results of  
379  $p$ -value inflation presented above and agree well with what is known for GWA

380 studies in general (Table 1). The main findings are as follows. In the presence of  
381 strong population structure and when relatedness was not accounted for, all  
382 tests displayed large numbers of false positives. When populations were  
383 simulated under random mating, the mean number of false positives was still  
384 large and exceeded the mean number of significant causal loci. In contrast, false  
385 positives were close to zero in all tests when accounting for relatedness and  
386 using GC to correct for any residual  $p$ -value inflation. The power to detect causal  
387 loci was always lower for binary GWA analyses compared to quantitative GWA  
388 analyses. Power to detect causal loci when accounting for relatedness as well as  
389 performing GC was generally low and decreased with increasing number of  
390 causal loci. For instance, with eight causal loci significant causal loci (one or  
391 more) could only be detected in 17 out of 300 data sets (pooled over all levels of  
392 population structure).

393  $P$ -value inflation was closely associated with long range LD caused by  
394 population stratification. In our simulated data sets, both the median and median  
395 absolute deviation for LD increased with population structure, at both short and  
396 long range (Fig. 2). A marked difference between short and long range LD was  
397 seen in the 95 % quantiles, where LD increased more with increasing population  
398 structure at long range (Fig. 2). Furthermore, LD half decay distance increased  
399 with increasing population structure (1.68 cM, 1.87 cM and 2.57 cM for  $N_e m =$   
400  $N_e/2$ ,  $N_e m = 2$  and  $N_e m = 1$ , respectively). Linkage disequilibrium plotted against  
401 physical distance for all levels of population structure are shown in Fig. S2  
402 (Supporting Information).

403

404 *Do binary and quantitative GWA associate the same genomic regions with traits*  
405 *under selection?*

406 There was a strong correlation between  $-\log_{10} p$  from binary and quantitative  
407 GWA analyses across all data sets when tests were conducted on the same  
408 phenotypic trait (Fig. 3 A and C). These correlations were stronger when  
409 ignoring relatedness (Fig. 3 A) compared to when relatedness was accounted for  
410 (Fig. 3 C). The correlations generally increased with increasing  $\log_{10} \bar{\lambda}$  (Fig. 3 A  
411 and C). When ignoring relatedness, the increase in correlation depended on  
412 population structure (Fig. 3 A) but was independent of population structure  
413 when accounting for relatedness (Fig. 3 C). This demonstrates that the  
414 underlying population stratification causes similar and strong biases in test  
415 statistics from GWA analyses testing for allele frequency change before and after  
416 selection and quantitative GWA analyses directly testing for associations  
417 between genotypes and traits under selection.

418       When the phenotypic traits under selection were based on different  
419 independent sets of causal loci and relatedness was ignored, 75% (inflated by a  
420 factor of 13.5 compared to a uniform null-distribution) of all correlations  
421 between  $-\log_{10} p$  from quantitative and binary GWA analyses were significant  
422 (Fig. 3 B). This dropped to 56% (inflated by a factor of 7.30 compared to a  
423 uniform null-distribution) when relatedness was accounted for (Fig. 3 D). When  
424 ignoring relatedness, this correlation increased with  $\log_{10} \bar{\lambda}$  for data sets with  
425 moderate and strong population structure but not for data sets simulated under  
426 random mating (Fig. 3 B). However, when relatedness was accounted for,  
427 correlations no longer increased with  $\log_{10} \bar{\lambda}$  for any level of population  
428 structure. Thus, even when variation in phenotypic traits was explained by

429 independent sets of loci in the binary and quantitative GWA analyses, the  
430 underlying population stratification caused  $p$ -values from these two tests to be  
431 similarly biased.

432

#### 433 *Allele frequency change in artificially selected house sparrow populations*

434 When testing for allele frequency change using gene-dropping simulations  
435 without GC, we found  $p$ -value inflation for both house sparrow populations (Fig.  
436 4; Leka:  $\lambda = 1.4$ ,  $SE=4.6 \times 10^4$ ; Vega:  $\lambda = 1.1$ ,  $SE=4.9 \times 10^4$ ). Without GC, The  
437 proportions of rejected null-hypotheses were estimated to 23 % at Leka and 9.4  
438 % at Vega. Furthermore, 33 loci were significant at  $q < 0.1$  in the Leka  
439 population, while no loci were significant (i.e. had  $q < 0.1$ ) in the Vega  
440 population. With GC,  $q$ -values for the most significant loci increased from 0.053  
441 to 0.51 at Leka and from 0.19 to 0.49 at Vega, and proportions of rejected null-  
442 hypotheses dropped to zero in both populations. Hence, after GC no loci showed  
443 larger allele frequency change than could be expected by random genetic drift  
444 alone.

445         When ignoring relatedness,  $p$ -value inflation with quantitative GWAS on  
446 tarsus length, was high in both populations (Leka:  $\lambda = 1.9$ ,  $SE = 1.5 \times 10^3$ ; Vega:  $\lambda$   
447  $= 1.7$ ,  $SE = 1.4 \times 10^4$ ). After accounting for relatedness,  $\lambda$ 's were below one for  
448 both populations and the  $q$ -values for the most significant loci were 0.91 and  
449 0.97 at Leka and Vega, respectively. Hence, after accounting for relatedness, no  
450 loci were significantly associated with tarsus length.

451         After accounting for relatedness,  $-\log_{10} p$  from GWA analyses for tarsus  
452 length and within year allele frequency change summed over the whole selection  
453 experiment (as tested by gene-dropping simulations) were significantly

454 correlated (Leka:  $r_p = 0.29$ ,  $t = 22$ ,  $df = 5029$   $p < 0.001$ ; Vega:  $r_p = 0.36$ ,  $t = 28$ ,  $df =$   
455  $5173$ ,  $p < 0.001$ ), with even stronger correlations when ignoring relatedness  
456 (Leka:  $r_p = 0.52$ ,  $t = 43$ ,  $df = 5129$ ,  $p < 0.001$ ; Vega:  $r_p = 0.43$ ,  $t = 35$ ,  $df = 5173$ ,  $p <$   
457  $0.001$ ). This suggests that artificial selection on tarsus length has influenced  
458 within year allele frequency changes within both islands (but see Discussion).

459

## 460 **Discussion**

461 Test statistics for allele frequency change in experimental evolution and GWA  
462 studies both ultimately rely on associations between genotypes and phenotypes  
463 (Fig. 3). As such, we here show that test statistics for allele frequency change and  
464 standard GWA analyses are equally prone to  $p$ -value inflation (Fig. 1, 3 and 4 and  
465 Table 1). However, we also show that methods to assess the magnitude of  $p$ -  
466 value inflation and account for relatedness in GWA studies are also applicable for  
467 testing for significant allele frequency change in experimental evolution studies  
468 (Fig. 1 and Table 1). Two additional benefits of using previously developed GWA  
469 approaches to assess the significance of allele frequency change are reduced  
470 computational time (at least relative to previously used permutation tests) and  
471 the possibility to account for additional covariates, but this is not considered in  
472 the present paper.

473 In permutation tests probability estimates are subject to error due to  
474 sampling the population of possible permutations (Ojala & Garriga 2010),  
475 generating a trade-off between precision of the  $p$ -values and computational  
476 resources. Previous studies assessing the significance of allele frequency change  
477 before and after selection by permutation have relied on only 1000 replicates  
478 (Gompert & Buerkle 2011; Pespeni *et al.* 2013). The minimum  $p$ -values one can

479 attain from such tests is the inverse of the number of replicates (one-tailed  
480 tests), which has the potential to lead to misleading results when correcting for  
481 multiple testing (Phipson & Smyth 2010) and does not allow for proper  
482 estimation of  $p$ -value inflation. In contrast, current GWA methods are optimized  
483 for large data sets and in the present paper we have demonstrated that they can  
484 be used to assess the significance of allele frequency change by fitting a binary  
485 response variable e.g. present/absent after an episode of selection. This enables  
486 accurate  $p$ -values for association statistics to be estimated much faster.

487         In our empirical data set from artificial selection on tarsus length in house  
488 sparrows, we report substantial  $p$ -value inflation for within year allele frequency  
489 change ( $p$ -values were attained from null-distributions generated by gene-  
490 dropping simulations rather than binary GWA analyses). By ignoring this  $p$ -value  
491 inflation, a substantial proportion of our loci (23% at Leka and 9.4% at Vega)  
492 would have erroneously been thought to be (directly or indirectly) associated  
493 with causal variants underlying variation in tarsus length. While we could not  
494 directly account for relatedness when estimating the  $p$ -values we could still  
495 perform GC. In doing so the expected number of significant loci dropped to zero  
496 in both populations. Hence, we emphasize that when testing for significance of  
497 allele frequency change, even in complex experimental designs spanning  
498 multiple generations,  $p$ -value inflation is an important confounding factor that  
499 potentially can be addressed with GC.

500

#### 501 *Power to detect loci under selection in experimental evolution studies*

502 The power to detect causal loci in GWA studies is largely determined by the  
503 number of causal loci, the difference in phenotypic values between alternative

504 allelic variants, and the degree of heterozygosity (Martin & Jiggins 2001; Korte &  
505 Farlow 2013). From a statistical perspective, quantitative traits are preferred  
506 over binary (case/control) because they improve power to detect a genetic effect  
507 (Bush & Moore 2012). This is also reflected here where the power of binary GWA  
508 analyses testing for allele frequency change before and after selection was  
509 always lower than quantitative GWA analyses performed directly on the  
510 underlying phenotypic trait under selection (Table 1).

511         Our simulated data were designed to mimic artificial selection  
512 experiments, where the selected phenotype is known and precise cut-off values  
513 for truncated selection can be used. The only variation with respect to survival of  
514 a particular phenotype in our simulations was environmental, specifically  
515 determined by the heritability of the trait under selection. In contrast, in natural  
516 selection (experiments) the researcher has no control over individual survival.  
517 As natural selection is subject to stochasticity, this generates additional variation  
518 (on top of environmental) with respect to the survival of a particular phenotype.  
519 Thus, we predict that the power to detect causal loci from test statistics for allele  
520 frequency change under natural selection (experiments) to be even lower than  
521 shown here.

522

### 523 *Linkage disequilibrium*

524 False statistical associations between genotypes and phenotypes are ultimately  
525 caused by long range LD in both GWA studies (Korte & Farlow 2013) and  
526 experimental evolution (Tobler *et al.* 2014; Schlötterer *et al.* 2015). Many  
527 biological processes, in particular mating among relatives (at any level of the  
528 population hierarchy) initially increase LD between loci across the whole

529 genome (Charlesworth & Charlesworth 2010; Kemppainen *et al.* 2015).  
530 Nevertheless, independent segregation and assortment of chromosomes ensures  
531 along with recombination that LD typically extends only short physical distances  
532 within chromosomes in large natural populations at any given time  
533 (Charlesworth & Charlesworth 2010). However, the fact that decay of LD can  
534 only take place in the presence of recombination that requires mating between  
535 individuals is often overseen. Thus, when the study sample comprises  
536 individuals from different populations (that do not meet to potentially mate),  
537 admixture LD, that is completely independent of physical distance, is created  
538 that will not decay with time (Fig. 2 and Fig. S2, Supporting information;  
539 Charlesworth & Charlesworth 2010; Kemppainen *et al.* 2015). This is the type of  
540 LD that is present in our simulated data with moderate and strong population  
541 structure. However, even in panmictic populations LD can be strong between  
542 physically distant pairs of loci due to genetic drift, selection and other sampling  
543 effects (particularly if  $N_e$  is small or only a few individuals have been selected or  
544 sampled; Charlesworth & Charlesworth 2010). This is evident from our data sets  
545 simulated under random mating despite large effective population sizes  
546 ( $N_e=10000$ ). When ignoring relatedness, long range LD was sufficient to cause at  
547 least one false positive in 37% of the data sets (Table 1), and 82% of all tests  
548 showed strong  $p$ -value inflation in the binary GWA analyses testing for allele  
549 frequency change before and after selection (see also Fig. 1). This was most likely  
550 because even in such cases there is variation in relatedness between individuals  
551 (i.e. all individuals are not equally related, or unrelated, to each other), which  
552 cause some population stratification in the data that is not easily detected by  
553 common population genetic tools. In other words, even in studies where



554 individuals are randomly sampled from large and arguably panmictic  
555 populations,  $p$ -value inflation in test statistics for allele frequency change may  
556 still be present (see also Tobler *et al.* 2014 and Schlötterer *et al.* 2015).  
557  
558 *Population stratification has strong influence on test statistics for allele frequency*  
559 *change in experimental evolution studies*  
560 It has been suggested that candidate genes from experimental evolution can be  
561 validated by GWA studies (Tobler *et al.* 2014; Schlötterer *et al.* 2015). In our  
562 simulated data  $p$ -values from quantitative and binary GWA analyses were much  
563 more correlated than expected by chance, when tests were conducted on the  
564 same data set but when the phenotypes were based on different and  
565 independent sets of causal loci (Fig. 3 B). Thus, here the correlations were  
566 caused by the underlying LD structure due to population stratification in the data  
567 rather than due to real genetic correlations, and this also occurred in randomly  
568 mating populations. Accounting for relatedness in both the quantitative and  
569 binary GWA analyses alleviated this to some extent (Fig. 3 D). Nevertheless, in  
570 data sets simulated under random mating,  $p$ -values were still inflated by a factor  
571 of 7.3 (compared to a null-distribution of no effect) resulting in significant  $p$ -  
572 value correlations in 56% of the data sets (Fig. 3 D).

573         It has been argued that due to allele frequency variation and possible  
574 epistatic interactions “lack of replication does not necessarily indicate lack of an  
575 effect”, if these tests are performed on different data sets (Schlötterer *et al.*  
576 2015). It is clear that the null-distribution of no effect when comparing  $p$ -values  
577 from allele frequency change and GWA analyses (on the trait under selection)  
578 does not lead to a uniform distribution of  $p$ -values. Instead it depends on the

579 genetic architecture of the data and the underlying population stratification.  
580 These were known for our simulated data and thus the results in Fig. 3 (C and D)  
581 can be considered as empirical null-distributions for the results in Fig. 3 (A and  
582 B). When a null-distribution cannot be created, the safest way to remove  
583 confounding effects of population stratification when validating candidate loci  
584 under selection with GWA studies is indeed to perform these tests on data sets  
585 from two different populations.

586           In experimental evolution, it is usually argued that parallel allele  
587 frequency changes in replicated selection experiments are the signature of  
588 selection (Tobler *et al.* 2014; Gompert *et al.* 2014; Schlötterer *et al.* 2015).  
589 However, following the argumentation above, if individuals in replicated  
590 selection experiments are sampled from the same population, the same  
591 underlying population stratification is likely to be present also among the  
592 individuals in the replicated experiments. This, in turn, may cause correlated  
593 allele frequency changes due to relatedness (long range LD) rather than true  
594 associations between the loci and causal genetic variants affecting the trait.  
595 However, also here the different methods to assess and correct for  $p$ -value  
596 inflation developed for GWA studies can potentially be used. In addition, to  
597 increase independence between replicated experiments, individuals could be  
598 collected from different populations, with the caveat that different causal  
599 variants then may be responsible for the traits in these populations.

600           In our artificial selection experiment  $-\log_{10} p$  from GWA analyses on  
601 tarsus length and allele frequency change were strongly correlated, even after  
602 accounting for relatedness. Thus, the allele frequency changes we observed were  
603 most likely due to the artificial selection on tarsus length that we imposed on the

604 populations. However, due to population stratification and the lack of a proper  
605 null-distribution (as argued above) we cannot exclude completely the possibility  
606 that the correlations we observed were caused by the underlying population  
607 stratification rather than selection on casual genetic variants affecting tarsus  
608 length. However, the fact that considerable  $p$ -value inflation in test statistics for  
609 allele frequency change existed (in particular in Leka; Fig. 4) suggests that  
610 evolutionary change in a heritable trait (or traits) indeed had occurred (see  
611 Kvalnes et al., in review). Nevertheless, we could not determine if any of the loci  
612 were associated with these traits, except through long range LD caused by  
613 population stratification in the data.

614

615 *Biological consequences of selection in the presence of population stratification*

616 In GWA studies  $p$ -value inflation is predominantly a statistical issue, i.e. it may  
617 lead to false claims of association between loci and the trait of interest. However,  
618 it should be recognized that allele frequency change due to selection in stratified  
619 populations (that causes  $p$ -value inflation) could have biological implications as  
620 well. If individuals with higher survival rates or reproductive success are more  
621 closely related than expected by chance (i.e. fitness depends on a heritable trait),  
622 any alleles that are identical by decent among the selected individuals are likely  
623 to hitchhike to higher frequency along with any causal variants for that trait. In  
624 natural populations, the biological consequences of this can for instance be; (1)  
625 reduced  $N_e$  (regardless of any eventual change in the census population size [ $N_c$ ])  
626 and as a consequence increased drift and rates of population differentiation, (2)  
627 inbreeding, (3) maladaptation and (4) reduced evolutionary potential. Below we

628 provide biological examples for scenarios 1-3. Evidence for scenario (4) follows  
629 indirectly from point (1).

630 (1) Exceptionally fast population differentiation was detected between  
631 geographically proximate populations of trout (*Salmo trutta*) that had undergone  
632 rapid adaptation to heavy metal contamination, relative to pristine populations  
633 much further apart (Paris *et al.* 2015). A reduction in  $N_c$  with a corresponding  
634 reduction in  $N_e$  could alone explain the fast drift within these populations.  
635 However, due to very strong selection, it is likely that the amount of drift was  
636 stronger than what could have been predicted solely by the reduction in  $N_c$ .  
637 According to our findings, this is particularly likely if selection operated on a  
638 highly heritable trait (possibly controlled by few genes of large effect) and  
639 populations exhibited strong population stratification before selection.

640 (2) Strong selection on heritable traits can directly lead to inbreeding, as  
641 then by definition individuals in the subset of the population that survives are  
642 likely to be more related to each other than expected by chance. Support for this  
643 comes from a study on an insular population of song sparrows (*Melospiza*  
644 *melodia*; Keller *et al.* 2001). This study showed that survival following a severe  
645 storm was not only higher for individuals with long wings (selection) but also for  
646 individuals with high inbreeding coefficients. Indirect evidence for this comes  
647 also from our artificial selection experiments in house sparrows presented here.  
648 The very existence of  $p$ -value inflation for allele frequency change suggests that  
649 non-random sampling with respect to relatedness between individuals indeed  
650 had occurred. Inbreeding potentially leads to inbreeding depression via  
651 increased genetic load (Charlesworth & Willis 2009), so strong selection on

652 heritable traits may have severe immediate negative consequences for the  
653 survival of the affected populations.

654 (3) Selection on heritable traits can lead to maladaptation when sub-optimal  
655 genotypes hitchhike to higher frequency due to LD (including long range LD e.g.  
656 caused by relatedness among the selected individuals) with loci under selection.  
657 That artificial selection in small populations can lead to maladaptation is already  
658 well known in commercial breeding programs (Garland 2003). This can e.g.  
659 clearly be seen in dogs where selective breeding has led to accumulation of  
660 negative mutations causing high prevalence of diseases in certain breeds  
661 (Marsden *et al.* 2016).

662 In other words, the potential biological consequences of strong selection  
663 in natural populations may have more important implications for conservation  
664 management strategies than previously recognized. This is expected to be  
665 especially true for, but not limited to, populations with strong population  
666 stratification.

667

### 668 *Conclusions*

669 As proof of concept, we have shown with simulated data that test statistics for  
670 allele frequency change before and after selection behave similarly to those from  
671 GWA studies on quantitative traits. Thus, the approaches and methods already  
672 available for GWA studies to account for relatedness and correct for  $p$ -value  
673 inflation is available also to experimental evolution studies. We emphasize that  
674 for any test statistic that ultimately depends on associations between genotypes  
675 and phenotypes, the potential of  $p$ -value inflation has to be considered and  
676 properly dealt with. Here we provide two examples of how this can be done:

677 binary GWA analyses when including relatedness as a random effect, and  
678 genomic control. Importantly, our study also shows using both simulations and  
679 empirical data from an artificial selection experiment in two free-living bird  
680 populations, that allele frequencies in large parts of the genome may change  
681 when selection is acting on a heritable trait. These genetic changes are likely to  
682 have considerable and wider consequences for the eco-evolutionary dynamics of  
683 such populations in the immediate future.

684

### 685 **Acknowledgements**

686 We are grateful to the inhabitants and farmers on Leka and Vega, whose  
687 hospitality enabled us to collect the empirical data used in this study. We thank  
688 R. Dahl, O.R. Davidsen, F. Jørgensen, T. Kolaas, L.K. Larsen, A. Lorås, P.A.  
689 Martinussen, R. Moe, M. Mørkved, R. Rismark, B.G. Stokke, and K. Sørensen for  
690 help during the field work, and Å.A. Borg Pedersen, O.R. Davidsen, R. Røsbak, and  
691 K. Yttersian Sletta for help with the laboratory work. We also thank Rowan  
692 Barrett who originally suggested using binary GWA analyses to assess the  
693 significance of allele frequency change. The empirical research was carried out in  
694 accordance with permits from the Norwegian Animal Research Authority  
695 (permits S-2007/1482 and ID-4011) and the Ringing Centre at Stavanger  
696 Museum, Norway. The study was supported by grants from the Research Council  
697 of Norway, project number 221956, and Strategic University Program (SUP) in  
698 Conservation Biology, project number 204303. This work was also partly  
699 supported by the Research Council of Norway through its Centres of Excellence  
700 funding scheme, project number 223257.

701

702 **Data Accessibility**

703 Data and R-code available from the Dryad Digital Repository:

704 <http://dx.doi.org/10.5061/dryad.vg4fj>

705

706 **Author contributions**

707 PK, HJ, BES, THR, BR, IJH and TK designed the project. PK executed all analyses  
708 and simulations. THR and HJ did most of the fieldwork for the artificial selection  
709 experiment. IJH, AMB, SL, HJ and AH generated SNP and reference genome data  
710 for the artificial selection experiment. PK, BR and HJ wrote the paper and all  
711 authors contributed with comments on the manuscript.

712

713 **References**

- 714 Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for  
715 genome-wide association analysis. *Bioinformatics (Oxford, England)*, **23**,  
716 1294–1296.
- 717 Barrett RDH, Hoekstra HE (2011) Molecular spandrels: tests of adaptation at the  
718 genetic level. *Nature Reviews Genetics*, **12**, 767–780.
- 719 Brakefield PM (2003) Artificial selection and the development of ecologically  
720 relevant phenotypes. *Ecology*, **84**, 1661–1671.
- 721 Burke MK, Dunham JP, Shahrestani P *et al.* (2010) Genome-wide analysis of a  
722 long-term evolution experiment with *Drosophila*. *Nature*, **467**, 587–590.
- 723 Bush WS, Moore JH (2012) Chapter 11: Genome-wide association studies. *PLoS*  
724 *computational biology*, **8**, e1002822.
- 725 Charlesworth B, Charlesworth D (2010) *Elements of evolutionary genetics* | Clc.  
726 Roberts and Company Publishers, Greenwood Village, Colorado.
- 727 Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nature*  
728 *Reviews Genetics*, **10**, 783–796.
- 729 Chen W-M, Abecasis GR (2007) Family-based association tests for genomewide  
730 association scans. *American journal of human genetics*, **81**, 913–926.
- 731 Clayton DG, Walker NM, Smyth DJ *et al.* (2005) Population structure, differential  
732 bias and genomic control in a large-scale, case-control association study.  
733 *Nature genetics*, **37**, 1243–1246.
- 734 Conner JK (2003) Artificial selection: a powerful tool for ecologists. *Ecology*, **84**,  
735 1650–1660.
- 736 Dabney A, Storey JD (2014) qvalue: Q-value estimation for false discovery rate  
737 control. *R package version 1.43.0*.
- 738 Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*,  
739 **55**, 997–1004.

- 740 Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to  
741 genetic-based association studies. *Theoretical population biology*, **60**, 155–  
742 166.
- 743 Dudley JW, Lambert RJ, La Roche De IA (1977) Genetic Analysis of Crosses  
744 among Corn Strains Divergently Selected for Percent Oil and Protein1. *Crop*  
745 *Science*, **17**, 111.
- 746 Ellegren H, Sheldon BC (2008) Genetic basis of fitness differences in natural  
747 populations. *Nature*, **452**, 169–175.
- 748 Excoffier L, Foll M (2011) fastsimcoal: a continuous-time coalescent simulator of  
749 genomic diversity under arbitrarily complex evolutionary scenarios.  
750 *Bioinformatics (Oxford, England)*, **27**, 1332–1334.
- 751 Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013) Robust  
752 demographic inference from genomic and SNP data. *PLoS Genetics*, **9**,  
753 e1003905.
- 754 Falconer DS, Mackay TFC (1996) *Introduction to Quantitative Genetics*. Longman,  
755 Essex.
- 756 Fay MP, Kim H-J, Hachey M (2007) On Using Truncated Sequential Probability  
757 Ratio Test Boundaries for Monte Carlo Implementation of Hypothesis Tests.  
758 *Journal of Computational and Graphical Statistics*, **16**, 946–967.
- 759 Garland T (2003) *Selection experiments: an under-utilized tool in biomechanics*  
760 *and organismal biology* (IL Bels, J-P Gasc, A Casinos, Eds.). BIOS Scientific  
761 Publishers Ltd, Oxford, Oxford.
- 762 Gompert Z, Buerkle CA (2011) A hierarchical Bayesian model for next-generation  
763 population genomics. *Genetics*, **187**, 903–917.
- 764 Gompert Z, Comeault AA, Farkas TE *et al.* (2014) Experimental evidence for  
765 ecological selection on genome variation in the wild. *Ecology letters*, **17**, 369–  
766 379.
- 767 Gratten J, Pilkington JG, Brown EA *et al.* (2012) Selection and microevolution of  
768 coat pattern are cryptic in a wild population of sheep. *Molecular ecology*, **21**,  
769 2977–2990.
- 770 Gromko MH (1995) Unpredictability of Correlated Response to Selection:  
771 Pleiotropy and Sampling Interact. *Evolution*, **49**, 685.
- 772 Gromko MH, Briot A, Jensen SC, Fukui HH (1991) Selection on Copulation  
773 Duration in *Drosophila melanogaster*: Predictability of Direct Response  
774 Versus Unpredictability of Correlated Response. *Evolution*, **45**, 69.
- 775 Hagen IJ, Billing AM, Rønning B *et al.* (2013) The easy road to genome-wide  
776 medium density SNP screening in a non-model species: development and  
777 application of a 10 K SNP-chip for the house sparrow (*Passer domesticus*).  
778 *Molecular ecology resources*, **13**, 429–439.
- 779 Hao K, Di X, Cawley S (2007) LdCompare: rapid computation of single- and  
780 multiple-marker  $r^2$  and genetic coverage. *Bioinformatics (Oxford, England)*,  
781 **23**, 252–254.
- 782 Heidaritabar M, Vereijken A, Muir WM *et al.* (2014) Systematic differences in the  
783 response of genetic variation to pedigree and genome-based selection  
784 methods. *Heredity*, **113**, 503–513.
- 785 Hill WG, Caballero A (1992) Artificial selection experiments. *Annual Review of*  
786 *Ecology and Systematics*, **23**, 287–310.
- 787 Hill WG, Weir BS (1988) Variances and covariances of squared linkage  
788 disequilibria in finite populations. *Theoretical population biology*, **33**, 54–78.



- 879 Illingworth CJR, Parts L, Schiffels S, Liti G, Mustonen V (2012) Quantifying  
 790 selection acting on a complex trait using allele frequency time series data.  
 791 *Molecular Biology and Evolution*, **29**, 1187–1197.
- 792 Jensen H, Saether BE, Ringsby TH *et al.* (2003) Sexual variation in heritability  
 793 and genetic correlations of morphological traits in house sparrow (*Passer*  
 794 *domesticus*). *Journal of evolutionary biology*, **16**, 1296–1307.
- 795 Jensen H, Steinsland I, Ringsby TH, Saether B-E (2008) Evolutionary dynamics of  
 796 a sexual ornament in the house sparrow (*Passer domesticus*): the role of  
 797 indirect selection within and between sexes. *Evolution*, **62**, 1275–1293.
- 798 Keller LF, Jeffery KJ, Arcese P *et al.* (2001) Immigration and the ephemerality of a  
 799 natural population bottleneck: evidence from molecular markers.  
 800 *Proceedings. Biological sciences / The Royal Society*, **268**, 1387–1394.
- 801 Kempainen P, Knight CG, Sarma DK *et al.* (2015) Linkage disequilibrium  
 802 network analysis (LDna) gives a global view of chromosomal inversions,  
 803 local adaptation and geographic structure. *Molecular ecology resources*, **15**,  
 804 1031–1045.
- 805 Korte A, Farlow A (2013) The advantages and limitations of trait analysis with  
 806 GWAS: a review. *Plant methods*, **9**, 29.
- 807 Marchini J, Cardon LR, Phillips MS, Donnelly P (2004) The effects of human  
 808 population structure on large genetic association studies. *Nature genetics*,  
 809 **36**, 512–517.
- 810 Marsden CD, Ortega-Del Vecchyo D, O'Brien DP *et al.* (2016) Bottlenecks and  
 811 selective sweeps during domestication have increased deleterious genetic  
 812 variation in dogs. *Proceedings of the National Academy of Sciences*, **113**, 152–  
 813 157.
- 814 Martin SH, Jiggins CD (2001) *Genomic Studies of Adaptation in Natural*  
 815 *Populations*. John Wiley & Sons, Ltd, Chichester, UK.
- 816 McCarthy MI, Abecasis GR, Cardon LR *et al.* (2008) Genome-wide association  
 817 studies for complex traits: consensus, uncertainty and challenges. *Nature*  
 818 *Reviews Genetics*, **9**, 356–369.
- 819 Nielsen R (2005) Molecular Signatures of Natural Selection. *Annu Rev Genet*, **39**,  
 820 197–218.
- 821 Ojala M, Garriga GC (2010) Permutation Tests for Studying Classifier  
 822 Performance. *The Journal of Machine Learning Research*, **11**, 1833–1863.
- 823 Palmer JO, Dingle H (1986) Direct and Correlated Responses to Selection Among  
 824 Life-History Traits in Milkweed Bugs (*Oncopeltus fasciatus*). *Evolution*, **40**,  
 825 767.
- 826 Pardo-Diaz C, Salazar C, Jiggins CD (2015) Towards the identification of the loci  
 827 of adaptive evolution. *Methods in ecology and evolution / British Ecological*  
 828 *Society*, **6**, 445–464.
- 829 Paris JR, King RA, Stevens JR (2015) Human mining activity across the ages  
 830 determines the genetic structure of modern brown trout (*Salmo trutta* L.)  
 831 populations. *Evolutionary applications*, **8**, 573–585.
- 832 Parts L, Cubillos FA, Warringer J *et al.* (2011) Revealing the genetic structure of a trait by  
 833 sequencing a population under selection. *Genome research*, **21**, 1131–1138.
- 834 Pespeni MH, Sanford E, Gaylord B *et al.* (2013) Evolutionary change during  
 835 experimental ocean acidification. *Proceedings of the National Academy of*  
 836 *Sciences*, **110**, 6937–6942.
- 837 Phipson B, Smyth GK (2010) Permutation P-values should never be zero:

838 calculating exact P-values when permutations are randomly drawn.  
839 *Statistical applications in genetics and molecular biology*, **9**, Article39.  
840 Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population  
841 stratification in genome-wide association studies. *Nature Reviews Genetics*,  
842 **11**, 459–463.  
843 Remolina SC, Chang PL, Leips J, Nuzhdin SV, Hughes KA (2012) Genomic basis of  
844 aging and life-history evolution in *Drosophila melanogaster*. *Evolution*, **66**,  
845 3390–3403.  
846 Rønning B, Broggi J, Bech C, Moe B (2016) Is basal metabolic rate associated with  
847 recruit production and survival in free-living house sparrows? *Functional*  
848 *Ecology*, 1140–1148.  
849 Schlötterer C, Kofler R, Versace E, Tobler R, Franssen SU (2015) Combining  
850 experimental evolution with next-generation sequencing: a powerful tool to  
851 study adaptation from standing genetic variation. *Heredity*, **114**, 431–440.  
852 Tobler R, Franssen SU, Kofler R *et al.* (2014) Massive habitat-specific genomic  
853 response in *D. melanogaster* populations during experimental evolution in  
854 hot and cold environments. *Molecular Biology and Evolution*, **31**, 364–375.  
855 Turner TL, Miller PM (2012) Investigating natural variation in *Drosophila*  
856 courtship song by the evolve and resequence approach. *Genetics*, **191**, 633–  
857 642.  
858 Turner TL, Miller PM, Cochrane VA (2013) Combining genome-wide methods to  
859 investigate the genetic complexity of courtship song variation in *Drosophila*  
860 *melanogaster*. *Molecular Biology and Evolution*, **30**, 2113–2120.  
861 Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM (2011) Population-based  
862 resequencing of experimentally evolved populations reveals the genetic  
863 basis of body size variation in *Drosophila melanogaster*. *PLoS Genetics*, **7**,  
864 e1001336.  
865 Zhou D, Udpa N, Gersten M *et al.* (2011) Experimental selection of hypoxia-  
866 tolerant *Drosophila melanogaster*. *Proceedings of the National Academy of*  
867 *Sciences*, **108**, 2349–2354.  
868

869 **Table 1.** Number of false positives and power of binary and quantitative GWA analyses for simulated data with two, four or eight causal  
870 loci. “K” indicates whether relatedness was included as a random effect. “GC” indicates if genomic control was performed. “Population  
871 structure” indicates if data was simulated with (Yes:  $N_e m = 1$ , i.e. ‘strong population structure’) or without (No: data simulated under  
872 random mating) population structure. “False Positives” is the mean (SD) number of significant loci further than 50 kbp (~5 cM) away  
873 from the closest causal locus and “Power” is the mean (SD) number of significant causal loci. \* Indicates that the test is equivalent to  
874 permutation test for allele frequency change before and after selection, within a single generation. See text for details on the simulations.

---

876	K	GC	Population	Number of	<u>Binary GWA analyses</u>		<u>Quantitative GWA analyses</u>	
877			Structure	Causal Loci	False Positives	Power	False Positives	Power
878	No	No	No	2	5.08 (11.78)*	0.98 (0.71)*	18.59 (27.77)	1.73 (0.60)
879	No	No	Yes	2	893.12 (1349.58)*	1.35 (0.77)*	1303.49 (1462.13)	1.82 (0.46)
880	No	Yes	No	2	0.03 (0.22)	0.38 (0.53)	0.09 (0.38)	1.22 (0.66)
881	No	Yes	Yes	2	0.00	0.21 (0.41)	0.00	0.41 (0.62)
882	Yes	No	No	2	0.18 (0.63)	0.61 (0.62)	0.66 (1.44)	1.55 (0.64)
883	Yes	No	Yes	2	0.39 (1.80)	0.57 (0.62)	0.71 (1.92)	1.34 (0.70)

884	Yes	Yes	No	2	0.05 (0.26)	0.47 (0.56)	0.20 (0.65)	1.42 (0.65)
885	Yes	Yes	Yes	2	0.07 (0.36)	0.42 (0.55)	0.12 (0.66)	1.20 (0.71)
886	No	No	No	4	4.83 (11.44)*	0.77 (0.92)*	16.78 (24.58)	2.32 (0.96)
887	No	No	Yes	4	1024.03 (1290.4)*	1.96 (1.56)*	1388.47 (1436.51)	2.81 (1.23)
888	No	Yes	No	4	0.07 (0.46)	0.15 (0.41)	0.00	0.40 (0.67)
889	No	Yes	Yes	4	0.01 (0.10)	0.04 (0.20)	0.00	0.11 (0.37)
890	Yes	No	No	4	0.46 (1.90)	0.32 (0.55)	0.18 (0.59)	1.08 (0.92)
891	Yes	No	Yes	4	0.23 (0.97)	0.22 (0.48)	0.51 (1.67)	0.83 (0.99)
892	Yes	Yes	No	4	0.10 (0.44)	0.22 (0.44)	0.05 (0.30)	0.74 (0.80)
893	Yes	Yes	Yes	4	0.11 (0.67)	0.15 (0.39)	0.18 (0.93)	0.44 (0.70)
894	No	No	No	8	4.46 (13.19)*	0.47 (0.89)*	13.81 (26.10)	1.49 (1.48)
895	No	No	Yes	8	709.24 (1054.34)*	2.46 (2.88)*	1041.55 (1179.46)	3.85 (2.83)
896	No	Yes	No	8	0.00	0.06 (0.31)	0.14 (0.98)	0.13 (0.42)
897	No	Yes	Yes	8	0.00	0.02 (0.14)	0.01 (0.10)	0.09 (0.38)
898	Yes	No	No	8	0.04 (0.24)	0.13 (0.39)	0.29 (1.51)	0.34 (0.77)

899	Yes	No	Yes	8	0.01 (0.10)	0.05 (0.22)	0.18 (1.20)	0.24 (0.53)
900	Yes	Yes	No	8	0.01 (0.10)	0.08 (0.31)	0.05 (0.41)	0.20 (0.59)
901	Yes	Yes	Yes	8	0.01 (0.10)	0.04 (0.20)	0.11 (1.10)	0.17 (0.47)

---

902

903

904  
905 Figure 1. Violin plots for  $p$ -value inflation as estimated by  $\lambda$  for GWA analyses on  
906 viability and on the underlying quantitative trait. When testing for allele  
907 frequency change before and after selection, viability was treated as a binary  
908 response variable (Binary). GWA analyses were also performed on the  
909 underlying quantitative trait under selection (Quantitative). Analyses were  
910 performed for three levels of population structure (Random mating, Moderate  
911 and Strong) when the trait was heritable ( $h^2=0.5$ ) and when (A) accounting for  
912 relatedness or (B) ignoring relatedness (see text for details). Binary GWA  
913 analyses not accounting for relatedness (B) is here used as proxy for previously  
914 used permutation tests for testing allele frequency change before and after  
915 selection. The dashed lines indicate  $\lambda = 1.1$ , above which we here consider  $p$ -  
916 value inflation to be strong.

917

918 Figure 2. Summary statistics of LD from simulated data with different levels of  
919 population structure. Statistics are shown for pairwise values of  $r^2$  between 500  
920 randomly chosen loci from each simulated data set ( $n = 300$ ). Results are shown  
921 for LD at short range ( $< 10$  kbp,  $\sim 1$  cM) and between loci on different  
922 chromosomes as a proxy for all long range LD.

923

924 Figure 3. Correlation between  $-\log_{10} p$  from quantitative and binary GWA  
925 analyses depends on  $\log_{10} \bar{\lambda}$ . The  $t$ -statistic refers to correlations between  $-\log_{10}$   
926  $P$  values between quantitative and binary GWA analyses when based on the  
927 same (A, C) or different (B, D) sets of causal loci.  $\log_{10} \bar{\lambda}$  is the mean inflation  
928 factor for the  $-\log_{10} p$  -values from each of the tests. In the upper panel (A, B)

929 individuals' relatedness is not taken into account while in the lower panel (C, D)  
930 relatedness (IBS) between all pairs of individuals was included as a random  
931 effect. Correlation coefficients ( $r_p$ ) are given in the figure and colored according  
932 to degree of population structure. The  $r_p$  in black represents all data points  
933 pooled. All significant tests ( $\alpha=0.05$ ; indicated by \*) have  $p<0.01$ . The vertical  
934 dashed line indicates the  $t$ -statistic for significance level  $p = 0.05$  (with  $df = 4998$ )  
935 for the original correlation test between  $-\log_{10} p$  between binary and  
936 quantitative GWA analyses.

937

938 Figure 4. Q-Q plot for expected versus observed  $-\log_{10} p$  from within year allele  
939 frequency change due to artificial selection on tarsus length. The slopes for  
940 regression lines equals  $\lambda$ . The dashed black line indicates a line with slope = 1  
941 and intercept = 0, and is shown for comparison.